Ute Heinze

Web Content Mining for Analyzing Job Requirements in Online Job Advertisements

Edition Wissenschaft
Apprimus

Web Content Mining for Analyzing Job Requirements in Online Job Advertisements

Von der Fakultät für Maschinenwesen der
Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades einer
D o k t o r i n d e r N a t u r w i s s e n s c h a f t e n
genehmigte Dissertation

vorgelegt von

Ute Heinze

(geb. Heckel)

Berichter:

Univ.-Prof. Dr. rer. nat. Sabina Jeschke
Univ.-Prof. Dr.-Ing. Klaus Henning

Tag der mündlichen Prüfung: 08. Oktober 2015

Ute Heinze

Web Content Mining for Analyzing Job Requirements in Online Job Advertisements

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

Ute Heinze:

Web Content Mining for Analyzing Job Requirements in Online Job Advertisements

1. Auflage, 2015

Apprimus Verlag, Aachen, 2015
Wissenschaftsverlag des Instituts für Industriekommunikation und Fachmedien
an der RWTH Aachen
Steinbachstr. 25, 52074 Aachen
Internet: www.apprimus-verlag.de, E-Mail: info@apprimus-verlag.de

ISBN 978-3-86359-386-5

D 82 (Diss. RWTH Aachen University, 2015)

Acknowledgments

The thesis is the result of my research at the Institute of Information Management in Mechanical Engineering (IMA), the Center for Learning and Knowledge Management (ZLW) and the Associated Institute for Management Cybernetics e.V. (IfU) of RWTH Aachen University. I owe the successful completion of my thesis to a number of people whom I would like to thank.

First and foremost, I would like to express my gratitude towards my supervisor Professor Sabina Jeschke who encouraged me to think outside of the box. Moreover, I would like to thank the members of the examination board Professor Klaus Henning, Professor Wolfgang Schröder and Professor Volker Stich for an inspiring discussion in my doctoral examination.

I am very thankful to my colleagues who supported me throughout the whole research process: Anja Richert, René Vossen, Christian Tummel, Ursula Bach, Olga Fidel, Anna Claus, Dörte Wego, Tobias Athmer and Andreas Burgdorf.

A special thank goes to Claudia Jooß and Katharina Schuster who always encouraged me to keep on going and whom I truly enjoyed working with.

Last but not least, I would like to thank my husband Martin for his serenity that gave me the strength to finish my thesis. I am especially grateful to my daughter Helena who saved me from going mad.

Aachen, December 2015

Ute Heinze

Abstract

The analysis of job requirements is crucial for several stakeholders. Companies define current and future personnel needs through job requirements and job seekers demand current information on job requirements to plan their careers. So far, job requirements were analyzed by applying hypothesis-testing approaches from the social sciences such as questionnaires or structured interviews. Moreover, as large numbers of data on job requirements are available through online job advertisements, also data-driven approaches were used such as keyword searches by pre-defined keyword lists. However, all of these methods can bias the research results as the researchers' opinions directly influence the expected outcomes. Therefore, the thesis deals with developing and evaluating an approach that enables independent analyses of job requirements. Hence, the web content mining process is implemented as a hypothesis-generating approach and is derived from the methods of big data analytics and knowledge discovery in databases. It combines methods from data mining, web mining, and natural language processing, and is especially adapted to the automatic analysis of job requirements in large numbers of online job advertisements. The web content mining process is based on the analysis of n-grams, word co-occurrences and their relative context information in online job advertisements. The process is applicable when no prior information on job requirements is available. The results show that the web content mining process helps to discover information on job requirements that other approaches do not provide such as information on non-academic qualification requirements and professional experiences. Thus, it independently discovers patterns in data without requiring any pre-defined research hypotheses. It is a valuable research method that complies with the general quality criteria of objectivity, validity, and reliability. When combined with a matching and recommendation component, the web content mining process can be integrated into an overarching recruiting 4.0 framework. Such a framework uses several software components to support decision-making in the recruiting process based on big data analytics. Future research shall concentrate on adapting the web content mining process to real-time data and to define the interfaces that are necessary to build the recruiting 4.0 framework.

Zusammenfassung

Die Analyse von Job-Anforderungen ist für verschiedene Zielgruppen von hoher Bedeutung. Einerseits definieren Unternehmen ihre momentanen und zukünftigen Personalbedarfe in Job-Anforderungen. Andererseits benötigen Jobsuchende aktuelle Informationen zu Job-Anforderungen, um ihre Karrieren zu planen. Bisher wurden Job-Anforderungen mit Hilfe von hypothesen-testenden Verfahren aus den Sozialwissenschaften untersucht wie bspw. Fragebögen und strukturierten Interviews. Aber auch daten-gestützte Verfahren auf Basis großer Datensätze an Job-Anforderungen in online verfügbaren Stellenanzeigen wurden angewandt wie bspw. die Schlagwortsuche anhand vordefinierter Schlagwortlisten. Alle diese Verfahren laufen jedoch Gefahr, von den Erwartungen und Meinungen der beteiligten Forscher bezüglich zu erwartender Ergebnisse unbewusst beeinflusst zu sein. Daher widmet sich die vorliegende Dissertation der Entwicklung und Evaluation eines unabhängigen Verfahrens zur Analyse von Job-Anforderungen. Im Zuge dessen wurde der Web Content Mining Prozess als hypothesen-generierendes Verfahren entwickelt und von den Methoden der Big Data Analyse und dem Prozess des Knowledge Discovery in Databases abgeleitet. Er kombiniert Methoden aus den Bereichen Data Mining, Web Mining sowie Natural Language Processing und wurde speziell für die automatische Analyse von Job-Anforderungen in großen Datensätzen von online Stellenanzeigen adaptiert. Er basiert auf der Analyse von N-Grammen, Wortkookkurrenzen und deren relativer Kontextinformation in online Stellenanzeigen. Er wird dann angewandt, wenn vorab keine Informationen über potenzielle Job-Anforderungen zur Verfügung stehen. Die Ergebnisse bestätigen, dass der Web Content Mining Prozess andere Informationen zu Job-Anforderungen zu Tage fördert als herkömmliche Ansätze wie bspw. Informationen zu nicht-akademischen Anforderungsprofilen oder geforderten beruflichen Erfahrungen. Demnach analysiert er auf Basis einer großen Datenmenge unabhängig Muster in den Daten und benötigt dafür keine vordefinierten Forschungshypothesen. Der Web Content Mining Prozess stellt eine wertvolle Wissenschaftsmethode dar, die den Qualitätsanforderungen wissenschaftlicher Forschung an Objektivität, Validität und Reliabilität vollends entspricht. Indem der Web Content Mining Prozess mit einer Matching- und Empfehlungskomponente ergänzt wird, entsteht ein übergreifendes Recruiting 4.0 Framework. Ein solches Recruiting 4.0 Framework integriert verschiedene Softwarekomponenten zur Entscheidungsunterstützung im Recruiting-Prozess auf Basis von Big Data Analysen. Zukünftige Forschungsansätze sollten sich der Analyse von Echtzeitdaten und der Integration des Web Content Mining Prozesses in ein übergreifendes Recruiting 4.0 Framework widmen.

Contents

List of Abbreviations

ACM	:	Association for Computing Machinery
API	:	Application Programming Interface
approx.	:	Approximately
ASCII	:	American Standard Code for Information Interchange
CEO	:	Chief Executive Officer
CVET	:	Continuing Vocational Education and Training
DB	:	Database
dblp	:	Computer science bibliography (formerly database and logic programming bibliography site)
df	:	Degrees of freedom
DNS	:	Domain Name System
DOM tree	:	Document Object Model tree
ε_{j0}	:	Expected frequencies of terms according to null hypothesis
GPS	:	Global Positioning System
H_0	:	Null hypothesis
HR	:	Human Resources
HRIS	:	Human Resources Information Systems
HTML	:	Hypertext Markup Language
IAB	:	Institute of Employment Research (Institut für Arbeitsmarkt- und Berufsforschung)
IE	:	Information Extraction
IR	:	Information Retrieval
IT	:	Information Technology
k	:	Number of categories in chi-square statistics
KDD	:	Knowledge Discovery in Databases

LIS	:	Library and Information Science
N	:	Sample size
n_j	:	Observed frequencies of terms
NLP	:	Natural Language Processing
NoSQL	:	Not only SQL
OECD	:	Organisation for Economic Co-operation and Development
OLAP	:	Online Analytical Processing
ω	:	Estimator for the effect size
p	:	Probability
p.	:	Page
PDF	:	Portable Document Format
PIAAC	:	Programme for the International Assessment of Adult Competencies
RQ	:	Research question
s_{max}	:	Maximum sample size
s_{min}	:	Minimum sample size
SEO	:	Search Engine Optimization
SME	:	Small and Medium-Sized Enterprise
SQL	:	Structured Query Language
TF-IDF	:	Term Frequency-Inverse Document Frequency
URL	:	Unified Resource Locator
VDI	:	The Association of German Engineers (Verein Deutscher Ingenieure)
WCM	:	Web Content Mining
χ^2	:	Chi-square value
XML	:	Extensible Markup Language

1 Introduction

1.1 Motivation

The current working world is in transition. Megatrends such as demographic and technological change as well as globalization influence established work environments profoundly. Demographic change causes a shift in the age patterns in society and the workforce towards a surplus of older employees and a decreasing number of middle-aged and younger employees. Companies face the challenge of a decreasing workforce, especially for young talents, and prolonged working lives. The technological change is represented by increasing degrees of digitalization, virtualization, and flexibility of the working world. Digitalization and virtualization describe the increasing number of modern communication and information technologies that enter traditional work settings. They offer new possibilities of work scenarios that are independent of time and space and, thus, lead to more flexibility. This comes along with an increased speed of working processes as well as change processes in the entire working world. Globalization describes the worldwide market development in the trade for products and services on the one side and the merge of the society to a common system on the other. It results in a growing complexity of the working environment and an increasing degree of global networking (BITKOM, 2013, p. 5; Rump, 2011, p. 37).

As consequence, companies face tremendous challenges in recruiting an appropriate workforce that responds to the changes in the working world. Often, a skills mismatch occurs. According to the Survey of Adult Skills by the Organisation for Economic Co-operation and Development (OECD), that was performed among the adult population aged between 16 and 65 years of 24 European countries, 10% of employees are over-skilled, i.e. they are more skilled than required by their job, and 4% are under-skilled on the opposite side (Pellizzari and Fichen, 2013; Desjardins and Rubenson, 2011; Handel, 2003). Bauer and Gartner (2014) state that up to 45% of unemployment on the German labor market are caused by a mismatch between requirements of a position and the skills of potential candidates (Bauer and Gartner, 2014, p. 1). According to Weitzel et al. (2014) finding appropriate candidates will be a tough endeavor for 35.8% of all open positions in Germany. They detect shortages in skilled labor that will cause 5.8% of open positions being expected not to be filled at all (Weitzel et al., 2014, p. 5).

Therefore, companies have to be supported to deal better with the challenge of recruiting a sustainable workforce. Jeschke warns that some of the traditionally strong business sectors in Germany such as mechanical engineering risk to fall behind global competitors regarding their innovative potential as they fail to adapt to the changing working world appropriately (cf. Wolking, 2015). However, new recruiting approaches are underway that are subsumed as ‘recruiting 4.0’

strategies referring to recent developments in industry 4.0 (Jeffrey, 2011; Jeschke, 2014; Weitzel et al., 2014). It is assumed that new technologies will revolutionize recruiting in the same way they have revolutionized industry before and have caused a fourth industrial revolution. Big data analytics is considered to be one of these revolutionizing technologies (Richtel, 2013; The Big Data Landscape, 2014). Big data analytics is characterized by the three Vs: volume, velocity, and variety (Dumbill, 2012, p. 10). First, big data analytics deal with the analysis of extra-large volumes of data that exceed the processing capacity of conventional database systems (volume). The Internet can be considered one such extra-large volume of data. Second, big data analytics deal with the analysis of data that change at a growing speed (velocity). They provide techniques to either analyze fast-moving data in real-time through data streams or to store the data for later batch processing. Moreover, third, big data analytics deal with a variety of data sources such as text, images, raw sensor data, or videos. Big data analytics can extract these unstructured data and to add meaning to humans or other applications that require structured inputs (Dumbill, 2012, p. 11). Thus, big data analytics transform big data to smart data that are enriched by meaning and allow decision makers to make smart decisions (cf. BMWi, 2015).

Web content mining is considered to be one such big data analytics tool (cf. Agneeswaran, 2012, p. 8). It can be applied to large data volumes on the web such as online job advertisements to leverage information on job requirements, and thus, to create smart data for recruiting. Online job advertisements correspond to the three Vs of big data analytics in the following way. First, online job advertisements are available at virtually unlimited numbers and constitute an extra-large data volume. Second, content, style and intentions of online job advertisements change at high speed as they are published solely for short periods of time on the web. Moreover, third, online job advertisements are highly diverse as they contain various formats such as text, images, or videos (cf. Weitzel et al., 2013, p. 54; Sailer, 2009, p. 199). Thus, the application of web content mining for analyzing online job advertisements yields promising insights into job requirements.

The analysis of job requirements is relevant for a variety of different stakeholders. First and foremost, employers benefit from a steady flow of information of job requirements. They define their current and future needs of certain qualifications and competencies among their workforce in job requirements (Sailer, 2009, p. 36; Sanchez-Cuadrado et al., 2010). These publicly available data are a resource to drive real-time competitor analyses in the context of the global war for talents (Dummert et al., 2014, p. 38). The knowledge of job requirements helps employers to set their company apart from its competitors. More accurately defined job profiles or an elaborate incentive system to attract highly skilled workers are the results. Thus, qualitative human resource planning can be taken to a new level with the help of web content mining on online job advertisements (cf. Watzka, 2014, p. 9). Second, job seekers are faced with the growing and continuing challenge of lifelong learning.

Therefore, they need information on current and future job requirements of potential employers. As the employers work in a fast-changing business environment, their job requirements also change fast. Job seekers are obliged to keep track of these fast-changing job requirements when they want to persist successfully in the labor market (cf. Wolking, 2015; Bensberg, 2014). This is especially important in fast-changing professions such as engineering or information technology (Litecky et al., 2010, p. 83; Smith and Ali, 2014, p. 203). Web content mining provides techniques for fast data collection and analysis. Furthermore, online job advertisements have a predictive character for current and future job requirements (Sailer, 2009, p. 38). Thus, web content mining on online job advertisements perfectly caters to the information needs of job seekers. Besides employers and job seekers, also educational and governmental institutions benefit from quickly accessible information about job requirements. For educational institutions job requirements are a guideline to shape educational programs and curricula in order to increase the employability of their graduates (Aken, 2008; Bensberg, 2014, 2012; Ciolac et al., 2010; Harper, 2012; Sanchez-Cuadrado et al., 2010). For governmental institutions job requirements are used as an indicator for the analysis of labor market developments (Sailer, 2009, p. 15).

Hence, the method of web content mining potentially delivers valuable insights into job requirements for a variety of different stakeholders. It is, therefore, in the center of attention of the present research work. The application domain of job requirements is used to examine the potentials of web content mining for automatically analyzing large data volumes of online job advertisements.

1.2 Research Design and Research Questions

The research design of the thesis is composed of the following steps. First, common methods for analyzing job requirements are characterized and compared. The comparison comprises not only methods of big data analytics but also quantitative methods of social research that are also applied for analyzing job requirements. The comparison aims to identify how job requirements have been analyzed so far and what methods yield what kind of results. By these data, the potentials for improvement offered by web content mining are derived. Second, the web content mining process for analyzing job requirements is designed, implemented, and applied. The application domain of the method of web content mining is German online job advertisements from two disciplines: engineering and business & economics. The focus was set to German online job advertisements in these two disciplines because they are especially affected by a fierce war for talents. This leads to a particularly high demand by employers and job seekers for fast and reliable information on the current labor market situation (cf. Bechmann et al., 2014; BMAS, 2014; Bundesagentur für Arbeit, 2014). Engineering related online job advertisements were chosen because the engineering sector is especially affected

by fast-changing job requirements due to fast-changing technologies. Thus, quick analyses of change processes in engineering are highly important for technology companies (cf. Lackmann and Koppel, 2015). The analysis of engineering related online job advertisements is contrasted by online job advertisements from business & economics. They were chosen because the business & economics sector disposes of the highest number of vacancies and, thus, the highest number of online job advertisements (cf. Müller et al., 2013). It ensures a sufficiently large data volume to demonstrate the applicability of web content mining. Furthermore, the online job adverts from both disciplines are used to demonstrate that the method of web content mining is not restricted to any specific application domain as most of the other methods are. In the next step, the results on job requirements that were gathered through web content mining are compared to the results of surveys in the field that used the common methods. This allows to determine as to whether the method of web content mining is capable of delivering at least equal or even better results than the other methods. Finally, the method of web content mining is evaluated. Therefore, the quality of the web content mining process is characterized according to its objectivity, validity, and reliability.

The following research questions (RQ) are asked and answered.

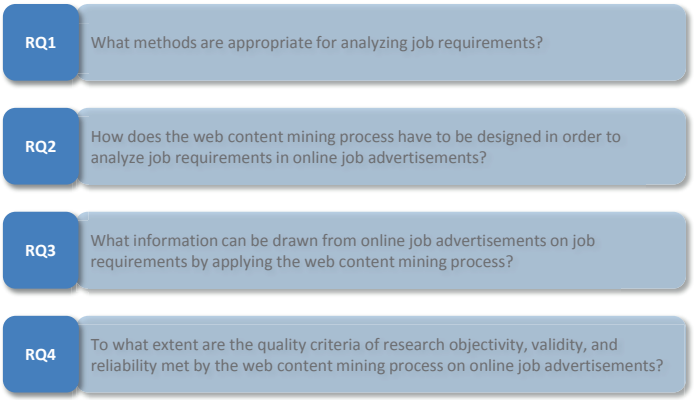

Figure 1: Research questions of the thesis

The following section gives an overview of the overall structure of the thesis and shows which chapters contain the answers to each of the research questions.

1.3 Thesis Outline

The thesis contains framework chapters and main chapters. The framework chapters introduce and conclude the research work. The main chapters lay the theoretical foundations of the research work and address the research questions subsequently. Figure 2 illustrates the general thesis outline.

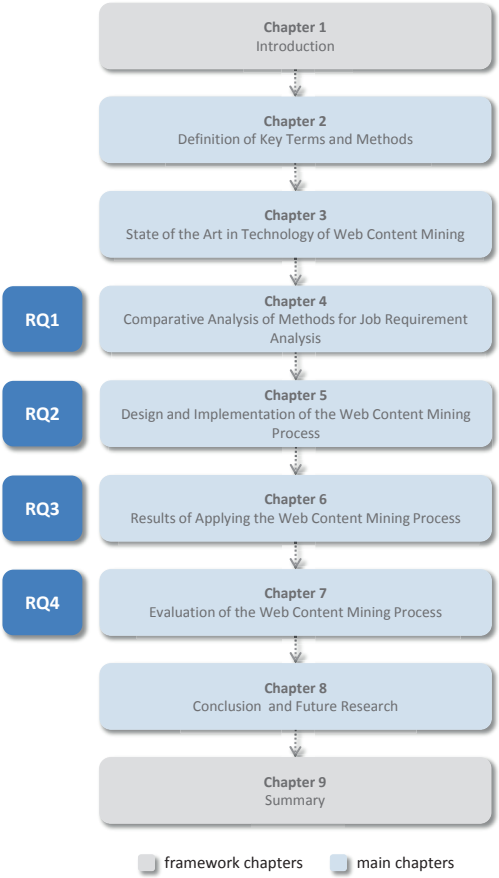

Figure 2: Thesis outline

Chapter 1 comprises the introduction and describes the motivation and the structure of the research project. Chapters 2 and 3 describe the theoretical and methodological background of the thesis. Chapter 2 deals with the definition of key terms that are related to the application domain of job requirements and online job advertisements. Furthermore, it introduces the theoretical background of quantitative methods of social research and job advertisement analysis that have been used for analyzing job requirements so far. Chapter 3 is devoted to the central method of the thesis: web content mining. It describes the current state of the art in research and technology of the method of web content mining and aims to define all necessary terms that are used throughout the thesis. Chapter 4 deals with the comparative analysis of previous surveys on job requirements. It aims to carve out what methods are appropriate to examine job requirements and, thus, gives a preliminary answer

to research question 1. The potentials for improvement offered by web content mining are derived and the functionality of the web content mining process is determined based on these results. A final judgment on the appropriateness of the method of web content mining for analyzing job requirements in online job advertisements is given in chapter 6. Chapter 5 is the key chapter of the thesis and describes the design and the implementation of the web content mining process for analyzing job requirements in online job advertisements. First, the design of the web content mining process is derived from the overarching process of Knowledge Discovery in Databases (KDD) which represents the answer to research question 2. Moreover, second, the implementation of each step in the web content mining process is described in detail. Chapter 6 presents the results for job requirements that were achieved by applying web content mining to online job advertisements. The chapter aims to show what information about job requirements can be drawn from online job adverts by using the web content mining process as developed throughout chapter 5. Moreover, it puts the gathered results into relation to the results of previous surveys on job requirements as introduced in chapter 4. It, thus, gives the answer to research question 3 and delivers judgment on the appropriateness of the method in addition to research question 1. Chapter 7 evaluates the applied method of web content mining. It critically reflects the quality of the method by the quality criteria of research objectivity, validity, and reliability. It, thereby, answers research question 4. Finally, chapter 8 draws general conclusions and describes the limitations of the method, the transfer of the results as well as future research perspectives. Chapter 9 finishes the thesis with a general summary of all results.